BEYOND THE GAMMA-POISSON

Improving the performance of Bayesian MCMC models for zero-inflated, overdispersed count data

Denwood, M.J.[1], Reid, S.W.J.[1], Toft, N.[2] and Innocent, G.T.[1]

m.denwood@vet.gla.ac.uk

[1] Comparative Epidemiology and Informatics, Institute for Comparitive Medicine, Department of Animal Production and Public Health, University of Glasgow Veterinary School, Bearsden Road, Glasgow G61 1QH, UK.

[2] Danish Meat Association, Vinkelvej 11, DK-8620 Kjellerup, Denmark

Introduction

Zero-inflated, overdispersed count data is usually analysed using a model based on the zero-inflated gamma Poisson (ZIGP) distribution. The ZIGP model assumes a mixture model of infected and uninfected groups, combined with a gamma distribution of means and Poisson distributions for each count in the infected group. This distribution is typically chosen for empirical fit, but using a lognormal distribution to describe the means would be more easily justified on the basis of biological assumptions if the distribution of means is thought to arise from a multiplicative process.

Materials and Methods

One hundred datasets with a combination matrix of 10 values of mean count, 5 values of variance:mean ratio, and zero-inflation of 0% and 20% were simulated using a Poisson distribution with both gamma and lognormal distributed means, for sample sizes of 20 and 200 counts. Each of the 400 datasets was then analysed using ZIGP and zero-inflated lognormal Poisson (ZILP) models implemented in Bayesian MCMC using the R package 'Bayescount' freely available from:

http://cran.r-project.org/web/packages/bayescount/

Results

A paired Wilcoxon test indicated that there was no significant difference between the median log-likelihoods of the ZIGP and ZILP models for the sample size 20 lognormal ($p=0.65$) or gamma ($p=0.37$) data, however at a sample size of 200 there was a location shift of +0.1% ($p<0.01$) for the ZIGP relative to the ZILP model with the gamma data, and -0.2% ($p<0.01$) for the lognormal data.

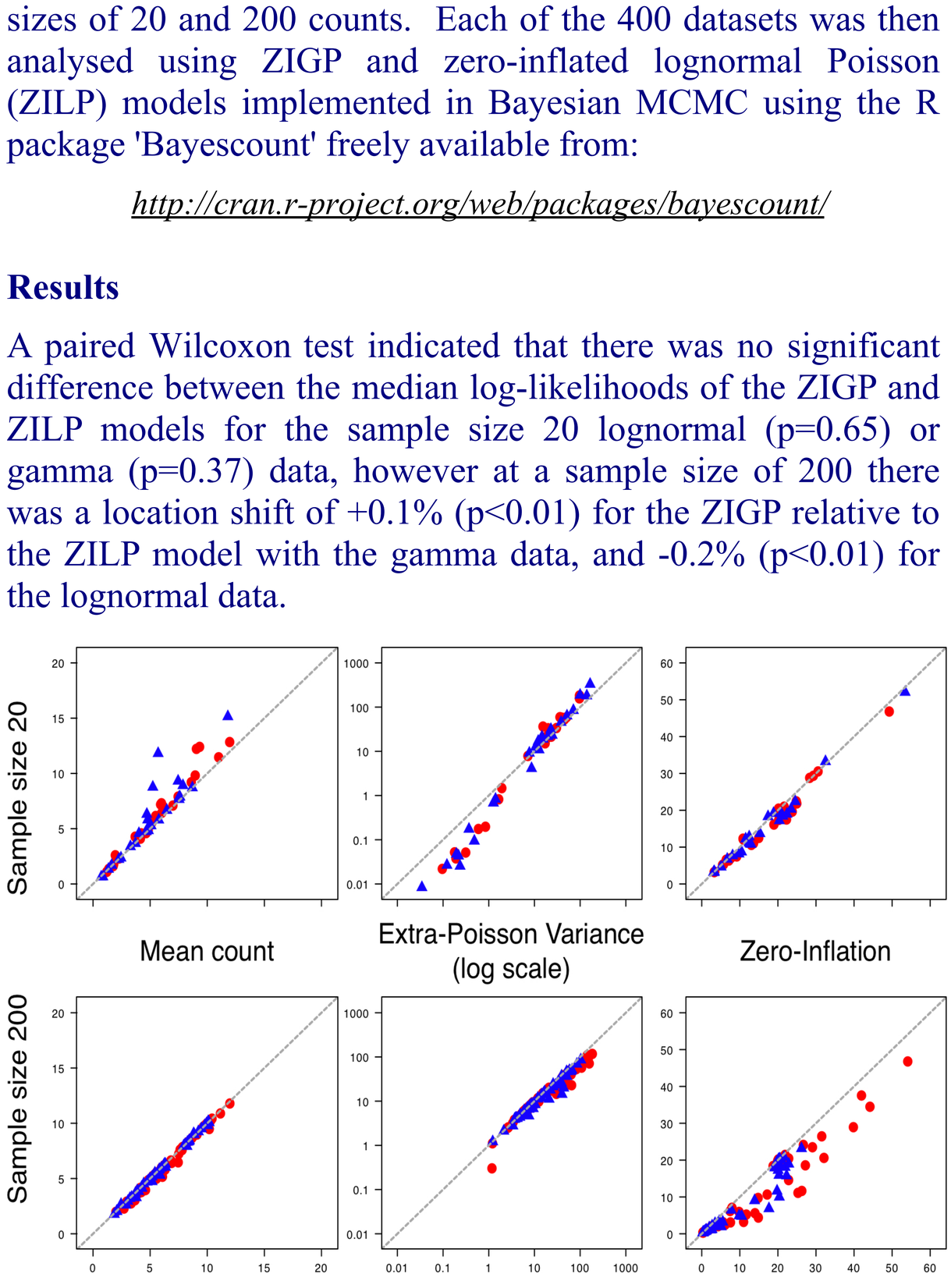

Figure 1: The ZILP (x axis) and ZIGP (y axis) median estimates for each parameter with lognormal (blue triangle) & gamma (red circle) simulated data

Dataset	ZILP Model		ZIGP Model	
	Mean time	SD	Mean time	SD
Lognormal size 20	6.13	0.34	8.30	1.22
Gamma size 20	6.04	0.21	8.39	1.23
Lognormal size 200	42.31	1.06	70.00	9.43
Gamma size 200	42.58	0.71	69.77	8.52

Table 1: The mean (standard deviation) time taken to complete each model in seconds, for each combination of model and sample size

The median estimates for each parameter were very similar between the models (*Figure 1*), although the zero-inflation estimates at sample size 200 were slightly higher on average for the ZILP model. Lower & upper credible intervals were also consistent between models (data not shown). The mean (sd) time taken to complete each simulation is shown in *Table 1*. The ZIGP model took on average 37% longer to analyse the data with a sample size of 20, and 65% longer with a sample size of 200, than the ZILP model.

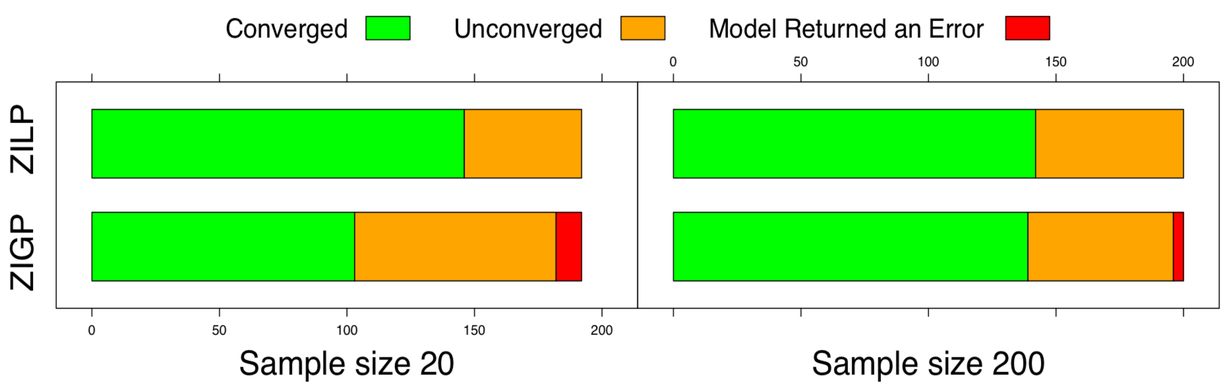

Figure 2: The number of datasets that were successfully and unsuccessfully analysed after 10000 iterations using each model at sample sizes of 20 and 200

The ZILP model was also able to achieve convergence (after 10000 iterations) for more datasets than the ZIGP model, with 43 more datasets (42%) successfully analysed by the ZILP model with a sample size of 20 (*Figure 2*).

Discussion

These results indicate that a ZILP model achieves similar results to a ZIGP model at the sample sizes tested. The comparison of log-likelihoods indicates that the two distributions are virtually indistinguishable at a sample size of 20, but that a small difference is appreciable at a sample size of 200. The ZILP model has the advantage of a biological justification if the factors contributing to the variation in mean count are thought to be multiplicative, for example through the combination of a series of probabilistic events. The improved performance of the ZILP model in terms of speed and convergence could also be of benefit, although the effect of the distribution choice at larger sample sizes requires further study.

In summary, the ZILP model is a practical and possibly more justifiable alternative to the ZIGP model, especially where model running time for larger datasets or improved convergence for smaller datasets is important.

Acknowledgments : This research was produced as part of the DEFRA-funded VTRI project 0101